



BANK OF CANADA
BANQUE DU CANADA

Technical Report No. 104 / Rapport technique n° 104

Variance Estimation for Survey-Weighted Data Using Bootstrap Resampling Methods: 2013 Methods-of-Payment Survey Questionnaire

by Heng Chen and Q. Rallye Shen



May 2015

**Variance Estimation for Survey-Weighted Data Using
Bootstrap Resampling Methods: 2013 Methods-of-
Payment Survey Questionnaire**

Heng Chen and Q. Rallye Shen

Currency Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
hchen@bankofcanada.ca
rshen@bankofcanada.ca

The views expressed in this report are solely those of the authors.
No responsibility for them should be attributed to the Bank of Canada.

Acknowledgements

We thank Geoffrey Dunbar, Shelley Edwards, Ben Fung, Kim P. Huynh, May Liu, Sasha Rozhnov and Kyle Vincent for their useful comments and encouragement. Maren Hansen provided excellent writing assistance. We also thank Statistics Canada for providing access to the 2011 National Household Survey and the 2012 Canadian Internet Usage Survey.

Abstract

Sampling units for the 2013 Methods-of-Payment Survey were selected through an approximate stratified random sampling design. To compensate for non-response and non-coverage, the observations are weighted through a raking procedure. The variance estimation of weighted estimates must take into account both the sampling design and the raking procedure. We propose using bootstrap resampling methods to estimate the variance. We find that the variance is smaller when estimated through the bootstrap resampling method than through Stata's linearization method, where the latter does not take into account the correlation between the variables used for weighting and the outcome variable of interest.

JEL classification: C83

Bank classification: Econometric and statistical methods

Résumé

Dans le cadre de l'Enquête sur les modes de paiement de 2013, les unités d'échantillonnage ont été sélectionnées à partir d'un plan d'échantillonnage stratifié aléatoire simple. Pour compenser la non-réponse et la non-couverture, les observations sont pondérées selon la procédure d'ajustement proportionnel itératif. L'évaluation de la variance des estimations pondérées doit tenir compte du plan d'échantillonnage et de la procédure d'ajustement proportionnel itératif. Les auteurs proposent d'utiliser la méthode de rééchantillonnage bootstrap pour évaluer la variance. L'étude montre que la variance estimée grâce à cette méthode est inférieure à celle obtenue avec la méthode de linéarisation propre au logiciel Stata, laquelle ne tient pas compte de la corrélation entre les variables utilisées pour la pondération et la variable de résultat étudiée.

Classification JEL : C83

Classification de la Banque : Méthodes économétriques et statistiques

1 Introduction

The Bank of Canada needs to understand and monitor Canadians' demand for cash. Because cash is an anonymous payment and it is difficult to obtain detailed characteristics of the cash users from the aggregate data, the Bank of Canada undertook the 2013 Methods-of-Payment (MOP) survey, which is a follow-up to the 2009 MOP (Arango and Welte 2012).

The 2013 MOP survey is designed to measure Canadian adult (over 18 years of age) consumers' attitudes toward and usage of different payment instruments, including cash, credit and newer methods such as the contactless feature of credit cards. A third party collected the data using an approximate stratified random sampling design. Henry, Huynh and Shen (2015) give an overview of the main results from the survey. Vincent (2015) constructs calibrated weights using the raking procedure.

This report serves as a technical companion to Henry, Huynh and Shen (2015) and proposes an approach to estimate the variances of the weighted means and proportions used in Henry, Huynh and Shen (2015). Variance estimates are crucial for building confidence intervals to assess the dispersion and for implementing statistical inferences to test various hypotheses.

In general, survey variance estimates depend on the specific weighting procedure, not just on the numerical values of the weights: variance estimates that disregard the weighting procedure are often biased. Hence an unbiased estimation method must incorporate two sources of randomness: (1) from the sampling design, which, in our case, is measured by the selection probability (the design weight) induced by stratified random sampling; and (2) from the raking procedure, which arises from adjusting the sample counts to match the population counts (calibrated weights).

If the sample units have full response and full coverage, we can simply use the inverse of the selection probability because our weights, and the variance estimates, will be straightforward to obtain. However, to adjust for non-response or non-coverage, additional modifications must be made to the weights through calibration. As a result, the final weights will depend on the particular calibration methods, such as non-response adjustment, post-stratification or raking. These methods will affect the variances of weighted estimates because calibrated weights are functions of the sample sizes (and not the population sizes) of the strata used for calibration, which are, in fact, random variables. If we ignore the calibration procedure, the confidence intervals will be conservative (Lu and Gelman 2003).

We propose using a bootstrap resampling method to take into account the effect of the raking procedure and to compute the variances of weighted estimates from the survey questionnaire (SQ) component of the 2013 MOP survey. We find that the variance is smaller when estimated with the bootstrap resampling method than with Stata's linearization method. Our

reason for choosing resampling over linearization is because the linearization method can be very difficult to implement due to the presence of nuisance parameters, such as the joint selection probability of two sampling units in the same stratum. The resampling method circumvents the requirement of explicitly evaluating the variance formula.

In Section 2, we discuss the different linearization variance estimates for design weights, post-stratification weights and raking weights. In Section 3, we choose our bootstrap resampling method among the various resampling methods. Further, instead of directly recomputing the statistics for each resample, we recompute the calibrated weights (replicate raking weights) for each resample.

In Section 4, we take a range of alternate variance estimates by methods such as linearization as benchmarks without specifying strata and linearization without correcting for the raking procedure. We do this so we can investigate the sources of discrepancies between different estimates. We focus on two variables: the *cash on hand* continuous variable and the *contactless* (*tap-and-go*) *credit usage* binary variable.

Section 5 concludes. In Appendix A, we provide a suggested workflow for generating replicate raking weights based on the raking procedure for Stata implementation. In Appendix B, we justify the raking procedure for the dual-frame survey design used in the 2013 MOP survey.

2 Variance Estimation Through Linearization

When discussing variance estimation of weighted total estimates, we follow the convention of denoting the finite population by $\mathcal{U}$ and the sample drawn from it by $\mathcal{S}$. For a given Canadian population $\mathcal{U}$ of individuals indexed by $i = 1, \dots, N$, we are interested in estimating the population total $T(Y)$ of a variable Y (e.g., if Y is a continuous variable for the value of an individual's *cash on hand*, then $T(Y)$ is the total value of *cash on hand* for the Canadian population. Similarly, if Y is a binary variable indicating usage of the contactless (*tap-and-go*) feature of a credit card, then $T(Y)$ is the total number of Canadians using the contactless credit feature of a credit card):

$$T(Y) \equiv \sum_{i \in \mathcal{U}} Y_i. \quad (1)$$

We are focusing on estimating the population total because many other population quantities of interest can be written in terms of $T(Y)$. For example, the population mean (e.g., the average value of a Canadian's cash holdings) can be written as $\bar{Y} \equiv \frac{T(Y)}{T(1)}$, where Y is the variable for the value of cash holdings and $T(1)$ is the population size for the 2013 MOP survey. Similarly, a subpopulation mean (e.g., the mean cash withdrawal of Canadian females) can be

written as $\bar{Y}_D \equiv \frac{T(ZY)}{T(Z)}$, where $Z = 1$ if an individual is female and $Z = 0$ otherwise.

Regardless of the weights used (either design or calibrated weights), the population total $T(Y)$ can be estimated as the total of the weighted sample variable y_j (here the sample quantities are denoted by lowercase letters):

$$t^A(y) \equiv \sum_{j \in S} w_j^A y_j, \quad (2)$$

where for individual j , w_j^A with $A \equiv \{D, PS, R\}$ denote different types of weights: w_j^D is the design weight, and w_j^{PS} and w_j^R are post-stratification and raking weights, respectively.

2.1 Design weights

The sampling design used for the 2013 MOP SQ approximates stratified random sampling. The third-party survey company determines the sample size of each stratum based on the expected response rate, and sampling units in each stratum are selected independently.

Let π_j be the selection probability for individual j , which is the probability that individual j is included in the sample. Assuming stratified random sampling, the Horvitz-Thompson estimator of $T(Y)$ is given by

$$\begin{aligned} t^D(y) &\equiv \sum_{j \in S} \frac{1}{\pi_j} y_j \\ &= \sum_{j \in S} w_j^D y_j, \end{aligned} \quad (3)$$

where the design weight $w_j^D = 1/\pi_j$ is the inverse of the probability individual j will be selected. Note that the w_j^D is fixed (non-random) and known before the survey is conducted. It is usually computed as the ratio between the census stratum count and the service-agreement targeted count.

The variance estimate of $t^D(y)$ is

$$\widehat{Var} [t^D(y)] \equiv \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}, \quad (4)$$

where π_{kl} is the probability of selecting the pair of sampling units (k, l) .

Due to non-response, however, the sample may have non-coverage (or under/over-coverage). We therefore need to calibrate the selection probabilities to account for non-response and non-

coverage. The weights are usually calibrated based on information from external sources.¹ For the 2013 MOP SQ, these external sources are the 2011 Census for the post-stratification and the 2012 Canadian Internet Use Survey (CIUS) for the raking. The variables that we select from the 2011 Census are gender, age and region, while the control variables from the 2012 CIUS in Vincent (2015) include the “Econ Plus” set of variables: marital status nested within region; age category nested within mobile phone ownership; age category nested within online purchase; income category nested within education, gender, home ownership; and employment status nested within region.

In the next subsection, we will discuss how to calibrate the design weights through either post-stratification or raking, which are both random variables, to compensate for non-response and non-coverage.

2.2 Calibrated weights

Calibrated weights are usually accomplished through two general methods: post-stratification and raking. Both methods are able to account for non-response² and non-coverage. In contrast to the non-random design weights, calibrated weights are random (Lu and Gelman 2003).

2.2.1 Post-stratification weights

Post-stratification adjusts the design weights to make the weighted data conform to the joint distribution of the post-strata variables in the external source. Creating post-stratification weights consists of breaking down the 2011 Census into post-stratification cells and then adjusting the design weights within each corresponding cell in the survey data (see Appendix A, Practical Implementation, Step 1). Specifically, for unit j in the post-stratification cell C_k , the post-stratification weight can be written as

$$w_j^{PS} = w_j^D \frac{\sum_{i \in \mathcal{U}} 1(i \in C_k)}{\sum_{l \in S} w_l^D 1(l \in C_k)}, \quad (5)$$

where $1(\cdot)$ is an indicator function taking the value 1 when its argument is true and 0 otherwise. Notice that the weights w_j^{PS} are random variables because they depend on the sizes of $C_k \cap S$,

¹ Another type of information, not from external sources, is the sample distributions of the variables. For example, using sample distributions of auxiliary variables from both respondents and non-respondents, logistic regression can be used to develop weights adjusted for non-response. Especially when a set of categorical variables is included, logistic regression weights are similar to raking weights (Kalton and Flores-Cervantes 2003). However, this alternative method is not pursued in the 2013 MOP SQ weighting.

² For the weight calibration to be successful in reducing non-response bias, the control variables need to be correlated with the response propensity or the outcome variables or both.

which are random. Then the weighted total estimator is

$$t^{PS}(y) \equiv \sum_{j \in S} w_j^{PS} y_j. \quad (6)$$

Kolenikov (2014) suggests that the estimated variance of $t^{PS}(y)$ is

$$\widehat{Var} [t^{PS}(y)] \equiv \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \mathbf{x}'_k \mathbf{b}^{PS}}{\pi_k} \frac{y_l - \mathbf{x}'_l \mathbf{b}^{PS}}{\pi_l}, \quad (7)$$

where

$$\mathbf{b}^{PS} \equiv \left(\sum_{j \in S} w_j^{PS} \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left(\sum_{j \in S} w_j^{PS} \mathbf{x}_j y_j \right). \quad (8)$$

The vector $\mathbf{b}^{PS}$ has the same dimension as $\mathbf{x}_j$ and

$$\mathbf{x}'_j \equiv (\delta_{1j}, \delta_{2j}, \dots, \delta_{Hj}),$$

where $\delta_{l,j} = 1$ if the sampling unit j is in the post-stratification cell l and 0 otherwise, and H is the number of post-stratification cells.

2.2.2 Raking weights

Since the post-stratification method adjusts every cell of a multi-way table, it can result in cells with zero or very small counts. In contrast, the raking method adjusts only the marginals, or the low-level interactions.³ Because there are many heuristic materials for the raking method, especially for its iterative proportional fitting algorithm, we omit the background introduction here. Readers interested in the raking will find more information in the 2013 MOP SQ weighting manual (Vincent 2015).

We frame the raking method as a one-step constrained minimization problem. The raking weights, $\{w_j^R\}_{j \in S}$, are computed by minimizing the objective function

$$\sum_{j \in S} [w_j \ln(w_j/w_j^D) - w_j + w_j^D], \quad (9a)$$

$$\text{such that } \sum_{j \in S} w_j \mathbf{x}_j = \mathbf{T}(\mathbf{X}), \quad (9b)$$

³ Low-level interactions are usually included by nesting two control variables; for example, nesting marital status within region (Vincent 2015).

where $T(\mathbf{X}) = \sum_{i \in \mathcal{U}} \mathbf{x}_i$ is the population total of control variables $\mathbf{X}$. The objective function is the discrepancy between the design weights w_j^D and raking weights w_j^R . On the right-hand side of the constraint are known or high-quality survey population marginal totals.⁴ See Appendix A, Practical Implementation, Step 2.

The resulting weighted total estimator is

$$t^R(y) \equiv \sum_{j \in \mathcal{S}} w_j^R y_j. \quad (10)$$

If the vector $\mathbf{X}$ of external control variables is used for raking, then Kolenikov (2014) suggests that the estimated variance of $t^R(y)$ is

$$\widehat{Var}[t^R(y)] \equiv \sum_{k,l \in \mathcal{S}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \mathbf{x}_k' \mathbf{b}^R}{\pi_k} \frac{y_l - \mathbf{x}_l' \mathbf{b}^R}{\pi_l}, \quad (11)$$

where

$$\mathbf{b}^R \equiv \left(\sum_{j \in \mathcal{S}} w_j^R \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \left(\sum_{j \in \mathcal{S}} w_j^R \mathbf{x}_j y_j \right). \quad (12)$$

The vector $\mathbf{b}^R$ has the same dimension as $\mathbf{x}_j$ and

$$\mathbf{x}_j' = \left(\delta_{1j}^{(1)}, \dots, \delta_{F_1j}^{(1)}, \delta_{1j}^{(2)}, \dots, \delta_{F_2j}^{(2)}, \dots, \delta_{1j}^{(H)}, \dots, \delta_{F_Hj}^{(H)} \right),$$

where $\delta_{kj}^{(l)} = 1$ if the sampling unit j is in the category k of the l -th control variable and 0 otherwise, F_l is the number of categories of the l -th control variable, and H is the number of external control variables $\mathbf{X}$.

Notice that the above expression is based on Deville and Särndal (1992), where they assumed $\sum_{j \in \mathcal{S}} w_j^D \mathbf{x}_j \approx T(\mathbf{X})$; that is, the estimator of $T(\mathbf{X})$, obtained by applying the design weights (w_j^D), is consistent. If this assumption is false, a more appropriate variance formula

⁴ Although the 2012 CIUS survey is calibrated to the 2011 Census, we do not consider the sampling variability of the population marginal totals from the CIUS because the sample size of the CIUS, at 22,615 individuals, is much larger than the 2013 MOP sample size. When such a large-scale survey is used for calibration, the sampling variability of the population marginal totals has a small impact and is usually ignored (Kolenikov 2014).

can be found in D'Arrigo and Skinner (2010), who rewrite

$$\begin{aligned} \text{Var} [t^R(y)] &= \text{Var} \left[\sum_{j \in S} w_j^R y_j \right] \\ &\approx \text{Var} \left[\sum_{j \in S} \frac{1}{\pi_j} [y_j - \mathbf{x}_j' \mathbf{b}^R] \right] \end{aligned} \quad (13a)$$

or

$$\approx \text{Var} \left[\sum_{j \in S} \frac{1}{\pi_j} \exp(\mathbf{x}_j' \hat{\boldsymbol{\lambda}}) [y_j - \mathbf{x}_j' \mathbf{b}^R] \right] \quad (13b)$$

where $\hat{\boldsymbol{\lambda}}$ are the Lagrange multipliers, which solve the constrained minimization problem. Hence the linearization variance estimator can be obtained by approximating $\text{Var} \left[\sum_{j \in S} w_j^R y_j \right]$ by $\text{Var} \sum_{j \in S} \left[\frac{1}{\pi_j} z_j \right]$ for a linearized variable z_j , where z_j is either $[y_j - \mathbf{x}_j' \mathbf{b}^R]$ or $\exp(\mathbf{x}_j' \hat{\boldsymbol{\lambda}}) [y_j - \mathbf{x}_j' \mathbf{b}^R]$.

Expression (13a) with $z_j = [y_j - \mathbf{x}_j' \mathbf{b}^R]$ is estimated by the $\widehat{\text{Var}} [t^R(y)]$ in equation (11), while expression (13b) is preferred under the non-response situation, when the assumption $\sum_{j \in S} w_j^D \mathbf{x}_j \approx \mathbf{T}(\mathbf{X})$ is less likely to hold. D'Arrigo and Skinner (2010) further demonstrate that these approximations are applicable to a stratified sample design with replacement sampling of clusters within strata.⁵

3 Variance Estimation Through Resampling

The above variance estimators that allow for calibrated weights are difficult to use in practice. First, these estimators use the selection probabilities π_j or, equivalently, the design weights w_j^D . Hence these estimators require the survey data set to include both calibrated weights and design weights, which may confuse users. Second, the end-user of the data set must be given the variables used for calibration, which may not be possible if confidential variables were used. Third, these estimators are not implemented in existing software packages. Finally, these variance estimators require joint selection probabilities π_{kl} , which are rarely computed in practice. Due to these complications, variance estimations of weighted totals usually proceed along the lines of resampling methods (Shao 1996, Kolenikov 2010).

There are three main resampling methods: balanced repeated replication (BRR), the jack-

⁵ Besides the "weighted residual" linearization approach described here, another possible approach is the hybrid linearization approach proposed by Lu and Gelman (2003), who decompose the variance conditioning on the sample size of post-strata and then linearize the variation of weights vector.

knife and the bootstrap (Rust and Rao 1996). Each involves creating multiple replicates of the data set by repeatedly sampling from the original sample. For example, the jackknife procedure for the stratified random design proceeds as follows: for stratum h , delete the k -th sampling unit and inflate the weights of the remaining units in stratum h to compensate for the deleted unit. Then perform the raking procedure on the remaining units and generate the hk -th replicate estimate $t^R(y)^{(hk)}$. Repeat the above steps for all numerations of h and k . The resulting jackknife variance estimate is

$$\widehat{Var}_{Jack} [t^R(y)] \equiv \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{k=1}^{n_h} \left\{ t^R(y)^{(hk)} - \widetilde{t^R(y)} \right\}^2 \quad (14)$$

where $\widetilde{t^R(y)} \equiv \sum_{h=1}^L \sum_{k=1}^{n_h} t^R(y)^{(hk)} / \sum_{h=1}^L n_h$, L is the number of strata and n_h is the number of sampling units in stratum h .

As for the bootstrap method, the parameter $t^R(y)^{(b)}$ is estimated in each b -th resampled data set using the same estimation procedure as for the original sample. The bootstrap variance estimator is then defined as

$$\widehat{Var}_{Boot} [t^R(y)] \equiv \frac{1}{B} \sum_{b=1}^B \left\{ t^R(y)^{(b)} - \overline{t^R(y)} \right\}^2 \quad (15)$$

where B is the number of bootstrap replicates and $\overline{t^R(y)} \equiv \sum_{b=1}^B t^R(y)^{(b)} / B$. To avoid the bias caused by small sample strata sizes n_h , we follow Rao and Wu's rescaling bootstrap, where scaled pseudo-values are internally generated (Rao and Wu 1988). The expression for these scaled pseudo-values can be found in Kolenikov (2010).

3.1 Choice of resampling method

We will focus on the bootstrap resampling method.⁶ We do not use BRR because it is more suitable for a stratified clustered sampling design, which was not used in our 2013 MOP SQ. The main reason that we choose the bootstrap over the jackknife is that the traditional delete-1 jackknife variance estimator will be inconsistent for non-smooth functions (e.g., sample quantiles). The consistent delete- d jackknife method requires a non-trivial specification for d , where there is a complicated interplay between the smoothness of the estimate and the parameter d .⁷

The bootstrap, on the other hand, will generally work for these non-smooth estimates, as discussed in Ghosh et al. (1984). Besides the major advantage of the bootstrap over the

⁶ Statistics Canada uses bootstrap procedures extensively. For example, the bootstrap replicate weights method is used in CIUS to estimate the coefficients of variation.

⁷ We thank May Liu for suggesting this point.

jackknife for non-smooth estimates, we prefer the bootstrap for two other reasons: (1) Computational burden: As pointed out in Kolenikov (2010), the replications of the delete- d jackknife increase notably with d , especially when applied to list-based establishment surveys; (2) Approximating distributions: The bootstrap can be used for estimating distributions and constructing more accurate one-sided confidence intervals, while the jackknife is typically only used for estimating variances.

Instead of recreating the sample in each replicate, we implement the more practical method of generating replicate weights using the *ipfraking* and *bsweights* commands in Stata. These replicate weights protect the privacy of sampling units and have the advantage of incorporating strata information as well as adjustments for non-response and non-coverage. The B sets of replicate raking weights are provided with the 2013 MOP SQ data set.

The construction of replicate raking weights under the bootstrap involves first taking the initial weights. Then a set of initial replication weights is constructed according to the bootstrap under stratified random sampling design. For example, if a unit from a replicate is not sampled, a zero weight is assigned to it and then the weights of other units in the same stratum are expanded to compensate. Next the raking method is applied to each of these sets of initial weights, generating the replicate raking weights. See Appendix A, Practical Implementation, Step 3.

In the report, the bootstrap sample size is set to be $n - 1$, in accordance with McCarthy and Snowden (1985) who propose this sample size for bootstrap with replacement. Furthermore, Kolenikov (2010) recommends the number of bootstrap replicates to be at least as large as the design degrees of freedom, so we can choose B to be 300. As shown in Figure A10 of Appendix A, estimated variance stabilizes after 300 bootstrap replications.

4 Results

Tables 2 to 6 show variance computations using different weights (design versus calibrated) and variance estimation methods (linearization versus resampling). We have calculated the variances of estimators for population totals, population means and the subpopulation means of various demographic strata.

The tables show results for the variables *cash on hand* and *tng_credit_year*. The first is the amount of cash the respondent has in his or her wallet, purse or pockets when completing the survey. Note that in Chen et al. (2014), the subpopulation means of *cash on hand* across different subsamples (online and offline) of the 2013 MOP SQ are computed and compared.

The second, *tng_credit_year*, is a binary variable indicating whether the respondent has used the contactless feature of a credit card in the past year. Taking the weighted total of this

variable, we find an estimate of the total number of people in the population that used the feature; taking the weighted mean, we obtain an estimate of the proportion of the population that has used it.

In all tables, the variances in the first three columns are calculated using Stata's linearization method, which does not take calibration into account. The variances in the last column are calculated by the bootstrap resampling method. Notice that there is a very small difference between the sample sizes for the two variables due to missing values, which will have a minor impact on imputation.

4.1 Comparison of the three linearization variance estimates

The differences among the estimated variances in the first three columns (those calculated by Stata's linearization method) are mainly due to the values of weights because the same formula is used in each case. From box plots of the weights displayed in Figure 1, it is clear that the raked weights are much more widely dispersed and contain more extreme values than the post-stratification weights. The same conclusion can be drawn from examining Table 1, which presents the 50th and 95th percentiles of the weights for various subsets of the sample. The table also shows that strata under-represented in the sample, such as the \$85K+ income category, are assigned higher weights by the raking procedure.

However, the effect of applying different weighting schemes also depends on the values of the variable being estimated, given that Stata computes the linearized variance estimate based on the weighted values of this variable (see notes to Tables 2 to 6). Figures 2 and 3 illustrate for cash on hand and contactless credit usage. The box plots in Figure 2 represent the distribution by gender of the weighted *cash on hand* values (the product of *cash on hand* and the weights). Comparing the design and post-stratification box plots, we see that *cash on hand* is more widely dispersed using post-stratification weights than design weights for females, while the opposite is true for males. Table 4 reflects this difference: The variance estimated using post-stratification weights is larger than that using design weights for females (1.29 times the design-weight variance) and smaller for males (0.90 times the design-weight variance). Furthermore, the plots for raked weights indicate that applying them increases the dispersion of weighted *cash on hand* for both males and females. This observation is again consistent with the results in Table 4 because the variance estimate increases to 1.65 for females and 1.37 for males.

Figure 3 contains box plots by gender of the weighted values for contactless credit usage (the product of *ing_credit_year* and the weights). The distributions for males and females are very similar, which is consistent with the male and female variance estimates in Table 4.

Furthermore, comparing Figures 2 and 3, it is apparent that applying the raked weights significantly increases the variation in *ing_credit_year* but not as much for *cash on hand*. This can be seen because the box plots for raked weights are much wider than the design-weighted box plots in Figure 3 but not in Figure 2. Upon examining the data, we find that respondents who have used the contactless feature in the past year, those having a value of 1 for *ing_credit_year*, are assigned very high raked weights. The 90th percentile of the raked weights for observations with *ing_credit_year* equal to 1 is about 15,500, for example, compared to only 9,872 for the post-stratification weights. These observations explain why, in Table 4, the variances for the raked-weighted estimators (column 3) are much higher than those for the design-weighted and post-stratification-weighted estimators (columns 1 and 2).

4.2 Comparison of the linearization and resampling variance estimates

When comparing the first column (design weights) and the fourth column (raked weights) in Tables 2 to 6, we observe that the ratios are sometimes larger than one. This observation does not contradict Deville and Särndal (1992), who claim that variances based on the raking procedure will be smaller than those based on design weights (see equations (4) and (11)) if $T(X) \approx \sum_{j \in S} w_j^D x_j$. As pointed out by Kalton and Flores-Cervantes (2003), when there is sizable non-response and non-coverage, the assumption of Deville and Särndal (1992) does not hold, and the estimates based on the raked weights will reduce bias at the expense of an increase in variance. Hence the ratios in the fourth column reflect the effect of non-response on the variance estimates.⁸

Next, when comparing the third and fourth columns in Tables 2 to 6, we observe that the resampling variance seems to be always smaller than the linearization variance based on the raked weights. At first glance this observation is puzzling: since the resampling variance captures the extra randomness from the weight adjustment procedure, this variance should be larger instead of smaller. The explanation is related to facts well known in the propensity score literature: The estimators based on the inverse of the non-parametric estimates of the propensity score, rather than on the true score, achieve the semi-parametric efficiency bound (Hirano, Imbens and Ridder 2003). In our current setting, we could say that the linearization variance estimate treats the raked weights as true weights (as fixed design weights), while the resampling variance estimate is using the information from the control variables in the raking procedure. With the raked weights, the variances calculated by resampling are smaller than those calculated by linearization.

⁸ In most payment surveys, variances are calculated by treating weights as fixed values (Angrisani, Foster and Hitzenko 2013).

4.3 Comparison of variance estimates with and without strata

Little difference exists between the variances calculated with and without strata, suggesting that the design strata variables (gender, region, and age) are not very strong predictors of cash on hand or contactless credit usage. Variances calculated with strata are commonly expected to be smaller than those calculated without, as is generally the case for population variances under proportional allocation (where strata sizes in the sample are proportional to those in the population). However, this result does not hold for variances under disproportional allocation. Further, it is always possible for a variance estimate from a particular sample to be larger under stratified sampling than it is under simple random sampling if the sample has large within-stratum sample variances (Lohr 2010).

For example, for some subpopulations, the estimated variances with post-stratification weights are larger with strata than without, as shown in Tables 2 to 6. The reason lies with the within-stratum sample variances and the values of the weights. It is noteworthy that sample variance estimates are being compared instead of population variances and that proportional allocation is not assumed with post-stratification weights. This is because the purpose of post-stratification is to correct for the over-representation of some strata and the under-representation of others. Therefore the variance after declaring strata could be higher because strata with higher within-stratum sample variances are weighted more heavily. A linear regression of the post-stratification weights against the within-stratum standard deviation of *cash on hand* shows that the weights and standard deviations do, in fact, exhibit a positive correlation, with the coefficient statistically significant at the 1 per cent level.

5 Summary

If variances for weighted estimates are computed without considering the raking procedure, the resulting confidence intervals will tend to be conservative. We therefore produce bootstrap replicate raking weights in Stata and use these to estimate the variances of weighted estimates from the 2013 MOP SQ.

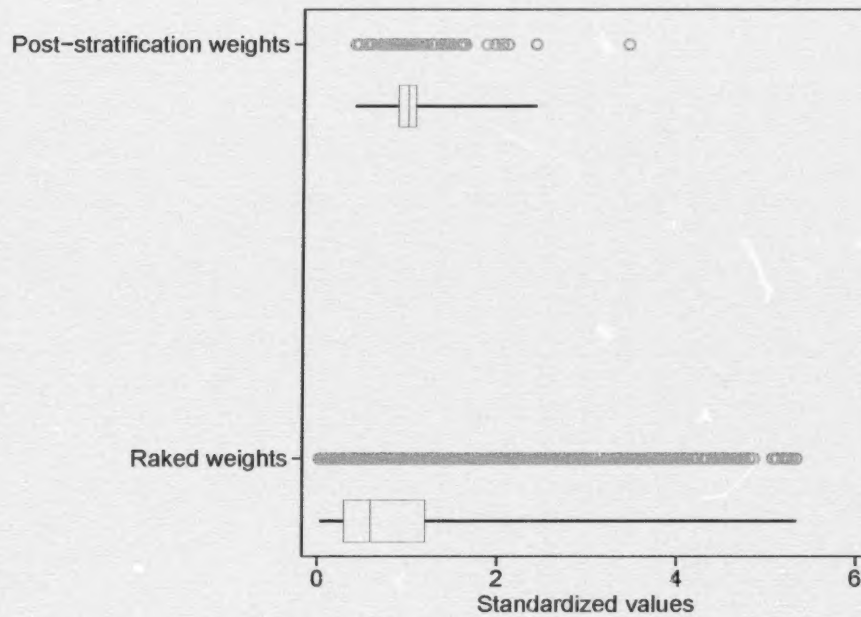
References

- Angrisan, M., K. Foster and M. Hitczenko. 2013. "The 2010 Survey of Consumer Payment Choice: Technical appendix."
- Arango, C. and A. Welte. 2012. "The Bank of Canada's 2009 Methods-of-Payment Survey: Methodology and Key Results." Bank of Canada Discussion Paper No. 2012-6.

- Brick, J. M., S. Dipko, S. Presser, C. Tucker and Y. Yuan. 2006. "Non-response Bias in a Dual Frame Sample of Cell and Landline Numbers." *Public Opinion Quarterly* 70 (5): 780–793.
- Callegaro, M., O. Ayhan, S. Gabler, S. Haeder and A. Villar. 2011. "Combining Landline and Mobile Phone Samples." *GESIS-Working Papers* .
- Chen, H., C. Henry, K. P. Huynh, Q. R. Shen and K. Vincent. 2014. "An Ounce of Planning is Worth a Pound of Weighting: Measuring Cash Holdings from the 2013 Bank of Canada Method-of-Payments Survey." Working Mimeo.
- D'Arrigo, J. and C. J. Skinner. 2010. "Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Non-response." *Survey Methodology* 36 (2): 181–192.
- Deville, J. C. and C. E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–382.
- Ghosh, M., W. Parr, K. Singh and G. J. Babu. 1984. "A Note on Bootstrapping the Sample Median." *The Annals of Statistics* 12 (3): 1130–1135.
- Henry, C., K. P. Huynh and Q. R. Shen. 2015. "2013 Methods-of-Payment Survey Results." Bank of Canada Discussion Paper No. 2015-4.
- Hirano, K., G. W. Imbens and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–1189.
- Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19 (2): 81–97.
- Kolenikov, S. 2010. "Resampling Variance Estimation for Complex Survey Data." *The Stata Journal* 10 (2): 165–199.
- . 2014. "Calibrating Survey Data Using Iterative Proportional Fitting (Raking)." *The Stata Journal* 14 (1): 22–59.
- Lohr, S. 2010. *Sampling: Design and Analysis*. Boston: Cengage Learning.
- Lu, H. and A. Gelman. 2003. "A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking." *Journal of Official Statistics* 19 (2): 133–151.
- Lumley, T. 2012. *Survey: Analysis of Complex Survey Samples*. R package version 3.28.2.

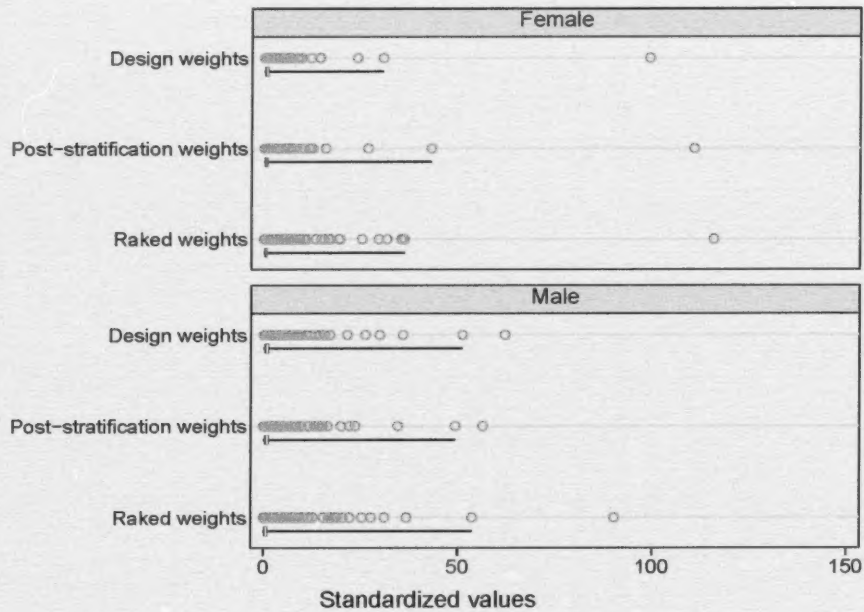
- McCarthy, P. J. and C. B. Snowden. 1985. "The Bootstrap and Finite Population Sampling." *Vital and Health Statistics* 95: 1–23.
- Quatember, A. 2016. *The Generation of Pseudo-Populations*. Heidelberg: Springer.
- Rao, J. N. K. and C. F. J. Wu. 1988. "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83 (401): 231–241.
- Roshwalb, A., N. El-Dash and C. Young. 2012. "Towards the Use of Bayesian Credibility Intervals in Online Survey Results." Ipsos-Reid: Public Affairs White Paper Series.
- Rust, K. F. and J. N. K. Rao. 1996. "Variance Estimation for Complex Surveys Using Replication Techniques." *Statistical Methods in Medical Research* 5: 283–310.
- Shao, J. 1996. "Resampling Methods in Sample Surveys (with Discussion)." *Statistics* 37: 203–254.
- StataCorp. 2013. *Stata 13 Base Reference Manual*. Stata Press, College Station, TX.
- Vincent, K. 2015. "2013 Methods-of-Payment Survey: Sample Calibration Analysis." Bank of Canada Technical Report No. 103.

Figure 1: Box plots of weights



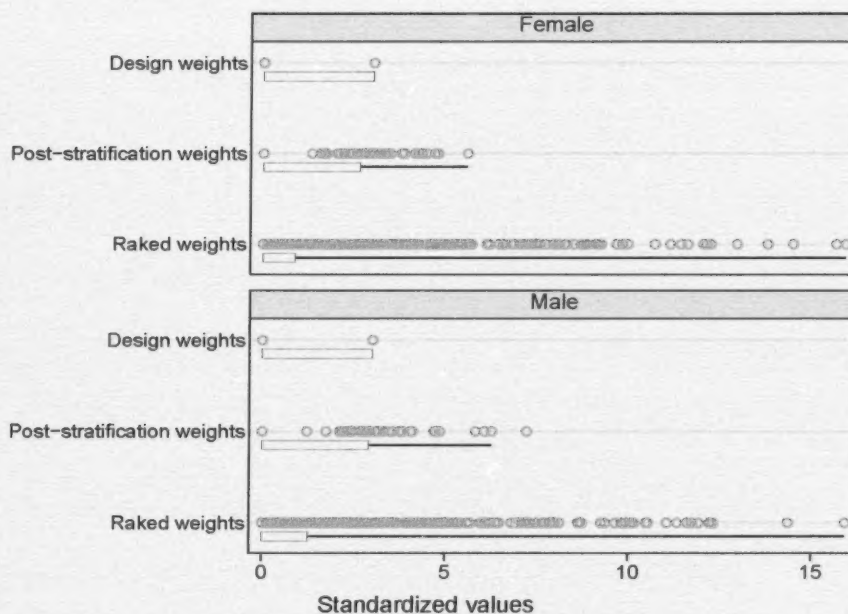
Note: For each set of weights, a dot plot of the values is shown above a box-and-whisker plot. The whiskers end at the 0.1 and 99.9 percentiles. The scale of the x-axis has been standardized by the mean of the raked weights.

Figure 2: Box plots of weighted *cash on hand* by gender



Note: The weighted values of *cash on hand* are shown. For each type of weight, a dot plot is shown above a box-and-whisker plot. The whiskers end at the 0.1 and 99.9 percentiles. The scale of the x-axis has been standardized by the mean of *cash on hand* weighted by the raked weights.

Figure 3: Box plots of weighted *tng_credit_year* by gender



Note: The weighted values of *tng_credit_year* (a binary variable) are shown. For each type of weight, a dot plot is shown above a box-and-whisker plot. The whiskers end at the 0.1 and 99.9 percentiles. The scale of the x-axis has been standardized by the mean of *tng_credit_year* weighted by the raked weights.

Table 1: 50th and 95th percentiles of weights

	Design Weights		Post-Stratification Weights		Raked Weights	
	50th	95th	50th	95th	50th	95th
Overall	7,235	7,235	7,128	10,395	4,221	25,169
Gender						
Female	7,235	7,235	7,128	10,395	4,088	23,616
Male	7,235	7,235	7,060	11,278	4,365	25,917
Age						
18-34	7,235	7,235	7,606	14,012	5,173	29,463
34-54	7,235	7,235	7,202	9,720	4,311	22,065
55+	7,235	7,235	7,060	9,207	3,520	25,917
Income						
<\$45K	7,235	7,235	7,128	10,197	3,165	21,040
\$45K-\$85K	7,235	7,235	7,060	10,395	3,917	24,152
\$85K+	7,235	7,235	7,202	10,700	7,293	38,298
Region						
Atlantic	7,235	7,235	7,202	10,700	7,293	38,298
Quebec	7,235	7,235	6,616	9,061	3,701	23,849
Ontario	7,235	7,235	6,464	10,395	3,789	22,198
Prairies	7,235	7,235	7,427	9,872	4,486	26,150
BC	7,235	7,235	7,764	13,407	4,622	29,420
Panel						
CFM	7,235	7,235	7,128	10,197	4,558	29,420
Non-CFM	7,235	7,235	7,128	10,395	4,814	24,180
Online	7,235	7,235	7,128	10,395	3,704	22,126

* The panel or subsample refers to the way the respondents were recruited (CFM: recently responded to the Canadian Financial Monitor (CFM) survey and invited by regular mail; Non-CFM: selected from an offline panel of potential volunteers and invited by regular mail; Online: selected from an online panel of potential volunteers and invited by email).

Note: The 50th and 95th percentiles of design, post-stratification and raked weights are shown. Raked weights are trimmed at five times their mean during the raking process.

Table 2: Total of *cash on hand* and *tng_credit_year*

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
<i>CASH ON HAND</i>				
<i>Total</i>	2.17E+09	1.00	1.02	1.02
Variance No Strata	9.79E+15	1.10	1.67	0.69
Variance Strata	9.67E+15	1.10	1.66	0.65
<i>TNG_CREDIT_YEAR</i>				
<i>Total</i>	9.28E+06	1.02	0.93	0.93
Variance No Strata	4.31E+10	1.12	2.51	1.07
Variance Strata	4.15E+10	1.09	2.54	1.14

* Linearization estimates are produced with the linearization procedure in Stata, which simply assumes a stratified random sample and does not take into account the weighting procedure. Hence all of the linearization variance estimates (for population totals) are calculated as $\sum_{h=1}^L (1-f_h) \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$, where L is the number of strata, n_h is the number of primary sampling units (PSUs) in stratum h , $(1-f_h)$ is the finite population correction for stratum h , f_h is the sampling rate for stratum h , y_{hi} is the weighted value for PSU (h, i) and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ (StataCorp 2013).

Note: All columns after the first have been divided by the value in the first column. The rows in italics show the weighted point estimates, and all other rows show variance estimates. For example, the total *cash on hand* is estimated to be 2.17E+09 under the design weights and 1.02*(2.17E+09) under the raked weights. The variance for this estimate is 9.67E+15 under the design weights with the linearization method, taking into account the strata, and 0.65*(9.67E+15) under raked weights with the resampling method, taking into account the strata. Total number of contactless credit adopters is estimated by the weighted total of the binary variable *tng_credit_year*. There are 3,663 respondents to the 2013 MOP survey, out of which 3,651 provided non-missing data for *cash on hand*. Among these 3,651 consumers, 1,743 are males and 1,908 are females. There are 3,578 respondents who provided non-missing data for *tng_credit_year*. Among these 3,578 consumers, 1,699 are males and 1,879 are females.

Table 3: Mean of *cash on hand* and *tng_credit_year*

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
<i>CASH ON HAND</i>				
<i>Mean</i>	81.96	1.00	1.03	1.03
Variance No Strata	14.02	1.09	1.51	0.85
Variance Strata	13.86	1.10	1.51	0.83
<i>TNG_CREDIT_YEAR</i>				
<i>Proportion</i>	0.36	1.02	0.93	0.93
Variance No Strata	6.43E-05	1.08	2.10	1.24
Variance Strata	6.20E-05	1.09	2.15	1.20

* Linearization estimates are produced with the linearization procedure in Stata, which simply assumes a stratified random sample and does not take into account the weighting procedure. Hence all of the linearization variance estimates (for population totals) are calculated as $\sum_{h=1}^L (1-f_h) \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$, where L is the number of strata, n_h is the number of PSUs in stratum h , $(1-f_h)$ is the finite population correction for stratum h , f_h is the sampling rate for stratum h , y_{hi} is the weighted value for PSU (h, i) , and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ (StataCorp 2013).

Note: All columns after the first have been divided by the value in the first column. The rows in italics show the weighted point estimates, and all other rows show variance estimates. The proportion of contactless credit adopters is estimated by the weighted mean of the binary variable *tng_credit_year*. There are 3,663 respondents to the 2013 MOP survey, out of which 3,651 provided non-missing data for *cash on hand*. Among these 3,651 consumers, 1,743 are males and 1,908 are females. There are 3,578 respondents who provided non-missing data for *tng_credit_year*. Among these 3,578 consumers, 1,699 are males and 1,879 are females.

Table 4: Mean of *cash on hand* and *tng_credit_year*, by demographic category

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post- Stratification Weights	Raked Weights	Raked Weights
CASH ON HAND				
GENDER				
<i>Female Mean</i>	69.53	1.02	1.10	1.10
Variance No Strata	26.16	1.29	1.63	0.96
Variance Strata	25.97	1.29	1.65	0.97
<i>Male Mean</i>	95.57	0.98	0.98	0.98
Variance No Strata	29.98	0.90	1.39	0.85
Variance Strata	29.68	0.90	1.37	0.94
AGE				
<i>18-34 Mean</i>	58.07	1.03	1.06	1.06
Variance No Strata	18.92	1.34	4.97	2.60
Variance Strata	19.01	1.34	4.96	2.61
<i>34-54 Mean</i>	68.54	1.01	1.02	1.02
Variance No Strata	11.38	1.04	1.68	1.26
Variance Strata	11.29	1.04	1.69	1.21
<i>55+ Mean</i>	112.91	1.00	1.05	1.05
Variance No Strata	83.81	1.12	1.00	0.71
Variance Strata	83.77	1.12	1.00	0.67
INCOME				
<i><\$45K Mean</i>	74.99	1.00	1.01	1.01
Variance No Strata	47.72	1.17	0.67	0.54
Variance Strata	47.51	1.17	0.67	0.53

Continued on next page

Table 4: Mean of *cash on hand* and *tng_credit_year*, by demographic category (continued)

	Linearization Estimates ^a			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
<i>\$45K-\$85K Mean</i>	89.50	1.00	1.06	1.06
Variance No Strata	36.29	1.01	2.55	1.43
Variance Strata	36.11	1.01	2.56	1.39
<i>\$85K+ Mean</i>	82.40	0.99	1.00	1.00
Variance No Strata	21.12	1.04	2.93	1.32
Variance Strata	20.94	1.05	2.93	1.55
<i>TNG_CREDIT_YEAR</i>				
GENDER				
<i>Female Proportion</i>	0.33	1.02	0.95	0.95
Variance No Strata	1.18E-04	1.08	2.16	1.36
Variance Strata	1.15E-04	1.09	2.20	1.45
<i>Male Proportion</i>	0.39	1.01	0.91	0.91
Variance No Strata	1.40E-04	1.07	2.05	1.23
Variance Strata	1.34E-04	1.09	2.11	1.42
AGE				
<i>18-34 Proportion</i>	0.40	1.00	0.88	0.88
Variance No Strata	2.63E-04	1.10	1.95	1.20
Variance Strata	2.54E-04	1.12	2.01	1.17
<i>34-54 Proportion</i>	0.35	1.02	0.99	0.99
Variance No Strata	1.64E-04	1.05	2.14	1.54
Variance Strata	1.57E-04	1.06	2.19	1.37
<i>55+ Proportion</i>	0.33	1.02	0.90	0.90
Variance No Strata	1.74E-04	1.07	2.13	1.33
Variance Strata	1.71E-04	1.08	2.17	1.36

Continued on next page

Table 4: Mean of *cash on hand* and *tng_credit_year*, by demographic category (continued)

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
INCOME				
<i><\$45K Proportion</i>	0.28	1.03	0.74	0.74
Variance No Strata	1.39E-04	1.09	1.65	1.21
Variance Strata	1.37E-04	1.09	1.66	1.24
<i>\$45K-\$85K Proportion</i>	0.37	1.01	0.96	0.96
Variance No Strata	1.77E-04	1.07	2.25	1.46
Variance Strata	1.74E-04	1.07	2.28	1.35
<i>\$85K+ Proportion</i>	0.48	1.01	0.91	0.91
Variance No Strata	3.12E-04	1.06	1.74	1.06
Variance Strata	3.09E-04	1.07	1.75	1.04

* Linearization estimates are produced with the linearization procedure in Stata, which simply assumes a stratified random sample and does not take into account the weighting procedure. Hence all of the linearization variance estimates (for population totals) are calculated as $\sum_{h=1}^L (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$, where L is the number of strata, n_h is the number of PSUs in stratum h , $(1 - f_h)$ is the finite population correction for stratum h , f_h is the sampling rate for stratum h , y_{hi} is the weighted value for PSU (h, i) , and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ (StataCorp 2013).

Note: All columns after the first have been divided by the value in the first column. The rows in italics show the weighted point estimates, and all other rows show variance estimates. The proportion of contactless credit adopters is estimated by the weighted mean of the binary variable *tng_credit_year*. There are 3,663 respondents to the 2013 MOP survey, out of which 3,651 provided non-missing data for *cash on hand*. Among these 3,651 consumers, 1,743 are males and 1,908 are females. There are 3,578 respondents who provided non-missing data for *tng_credit_year*. Among these 3,578 consumers, 1,699 are males and 1,879 are females.

Table 5: Mean of *cash on hand* and *tng_credit_year*, by region

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
CASH ON HAND				
Atlantic Mean	96.30	0.99	0.97	0.97
Variance No Strata	420.49	0.91	0.56	0.50
Variance Strata	418.01	0.92	0.56	0.53
Quebec Mean	85.47	1.02	0.89	0.89
Variance No Strata	100.11	1.35	0.31	0.34
Variance Strata	99.34	1.35	0.31	0.30
Ontario Mean	79.63	0.99	1.12	1.12
Variance No Strata	28.80	1.02	3.60	1.82
Variance Strata	28.38	1.02	3.62	1.84
Prairies Mean	80.74	1.00	1.08	1.08
Variance No Strata	21.81	1.28	2.74	1.76
Variance Strata	20.92	1.30	2.78	1.93
BC Mean	75.27	0.99	1.06	1.06
Variance No Strata	25.52	0.98	2.59	1.57
Variance Strata	24.86	0.99	2.61	1.41
TNG_CREDIT_YEAR				
Atlantic Proportion	0.33	0.97	0.85	0.85
Variance No Strata	8.42E-04	1.03	2.16	1.48
Variance Strata	8.51E-04	1.04	2.16	1.66
Quebec Proportion	0.26	1.03	0.92	0.92
Variance No Strata	2.11E-04	1.12	1.97	1.06
Variance Strata	2.04E-04	1.13	2.03	1.30

Continued on next page

Table 5: Mean of *cash on hand* and *tng_credit_year*, by region (continued)

	Linearization Estimates [*]			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
<i>Ontario Proportion</i>	<i>0.45</i>	<i>1.00</i>	<i>0.94</i>	<i>0.94</i>
Variance No Strata	1.87E-04	1.01	2.16	1.51
Variance Strata	1.85E-04	1.01	2.17	1.52
<i>Prairies Proportion</i>	<i>0.32</i>	<i>1.05</i>	<i>0.92</i>	<i>0.92</i>
Variance No Strata	3.73E-04	1.22	1.97	1.31
Variance Strata	3.72E-04	1.24	2.00	1.33
<i>BC Proportion</i>	<i>0.36</i>	<i>1.01</i>	<i>0.91</i>	<i>0.91</i>
Variance No Strata	4.66E-04	1.03	2.04	1.18
Variance Strata	4.63E-04	1.03	2.05	1.39

* Linearization estimates are produced with the linearization procedure in Stata, which simply assumes a stratified random sample and does not take into account the weighting procedure. Hence all of the linearization variance estimates (for population totals) are calculated as $\sum_{h=1}^L (1-f_h) \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$, where L is the number of strata, n_h is the number of PSUs in stratum h , $(1-f_h)$ is the finite population correction for stratum h , f_h is the sampling rate for stratum h , y_{hi} is the weighted value for PSU (h, i) , and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ (StataCorp 2013).

Note: All columns after the first have been divided by the value in the first column. The rows in italics show the weighted point estimates, and all other rows show variance estimates. The proportion of contactless credit adopters is estimated by the weighted mean of the binary variable *tng_credit_year*. There are 3,663 respondents to the 2013 MOP survey, out of which 3,651 provided non-missing data for *cash on hand*. Among these 3,651 consumers, 1,743 are males and 1,908 are females. There are 3,578 respondents who provided non-missing data for *tng_credit_year*. Among these 3,578 consumers, 1,699 are males and 1,879 are females.

Table 6: Mean of *cash on hand* and *tng_credit_year*, by subsample

	Linearization Estimates*			Resampling Estimates
	Design Weights	Post-Stratification Weights	Raked Weights	Raked Weights
CASH ON HAND				
CFM Mean	84.01	1.00	0.99	0.99
Variance No Strata	23.65	1.04	1.84	1.04
Variance Strata	23.46	1.04	1.84	1.10
Non-CFM Mean	92.15	1.02	1.00	1.00
Variance No Strata	150.44	1.21	0.39	0.38
Variance Strata	150.45	1.21	0.39	0.30
Online Mean	75.46	0.99	1.08	1.08
Variance No Strata	26.17	0.98	3.01	1.75
Variance Strata	26.04	0.98	3.03	1.85
TNG_CREDIT_YEAR				
CFM Proportion	0.39	1.01	0.89	0.89
Variance No Strata	1.81E-04	1.07	2.03	1.18
Variance Strata	1.79E-04	1.07	2.04	1.24
Non-CFM Proportion	0.35	1.03	0.95	0.95
Variance No Strata	3.27E-04	1.08	1.97	1.44
Variance Strata	3.24E-04	1.09	1.98	1.39
Online Proportion	0.33	1.02	0.95	0.95
Variance No Strata	1.42E-04	1.08	2.15	1.45
Variance Strata	1.40E-04	1.08	2.16	1.35

* Linearization estimates are produced with the linearization procedure in Stata, which simply assumes a stratified random sample and does not take into account the weighting procedure. Hence all of the linearization variance estimates (for population totals) are calculated as $\sum_{h=1}^L (1-f_h) \frac{n_h}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$, where L is the number of strata, n_h is the number of PSUs in stratum h , $(1-f_h)$ is the finite population correction for stratum h , f_h is the sampling rate for stratum h , y_{hi} is the weighted value for PSU (h, i) , and $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ (StataCorp 2013).

Note: All columns after the first have been divided by the value in the first column. The rows in italics show the weighted point estimates, and all other rows show variance estimates. The proportion of contactless credit adopters is estimated by the weighted mean of the binary variable *tng_credit_year*. There are 3,663 respondents to the 2013 MOP survey, out of which 3,651 provided non-missing data for *cash on hand*. Among these 3,651 consumers, 1,743 are males and 1,908 are females. There are 3,578 respondents who provided non-missing data for *tng_credit_year*. Among these 3,578 consumers, 1,699 are males and 1,879 are females. The panel or sub-sample refers to the way the respondents were recruited (CFM: recently responded to the Canadian Financial Monitor (CFM) survey and invited by regular mail; Non-CFM: selected from an offline panel of potential volunteers and invited by regular mail; Online: selected from an online panel of potential volunteers and invited by email).

Appendix A Practical Implementation

We implement variance estimation in Stata, a popular statistical software package. Referring to Kolenikov (2010, 2014), we use the *ipfraking* and *bsweights* commands. Stata do-files for replicating our process, and results are available upon request. In Table A1, we compare the *Rake* command in R with the *ipfraking* command in Stata so that similar results can be reproduced under R.

Step 1: Compute the initial post-stratification weight

First, we must specify the strata and the initial weight with which to begin the raking procedure. The raking procedure will minimize the discrepancy between the raked weights and initial weights. Here we set the initial weights to be the post-stratification weights *basewgt*, which equals the population stratum size divided by the number of respondents in the corresponding stratum. This *basewgt* variable will be declared in the argument [pw=varname] in the *ipfraking* command, which generates the raked weights.

The weights are also declared when setting the survey environment in Stata. For example, the Stata output in Figure A1 shows an estimate of the *cash on hand* variable using the design weights. To obtain this estimate in Stata, we declare the weight variable as well as the strata used in the sampling design, which divide the population based on region, gender, and age.

Figure A2 shows the command for declaring post-stratification weights in Stata. The post-strata variable, *strata*, contains identifiers for the strata to be used, and the postweight variable, *Count*, contains the population counts for each stratum. Declaring these two variables is equivalent to using *basewgt*, the variable containing the post-stratification weights.

Step 2: Generate the raked weights

Note that raked weights for the entire sample are necessary for generating replicate raking weights. Besides the initial post-stratification weights, the *ipfraking* command also requires us to specify the population marginal totals of the control variables. Hence, as mentioned above, we refer to the weighting manual and select the “Econ Plus” set of variables recommended in Vincent (2015), which are marital status nested within region; age category nested within mobile phone ownership; age category nested within online purchase; income category nested within education, gender, home ownership; and employment status nested within region.

We trim the weights at five times their mean to avoid extreme weights. Assuming the total calculated with the untrimmed weights to be the true value of total *cash on hand*, we compute the mean squared error (MSE) for weights trimmed at different values. Figure A3 shows the ratios of these MSEs to the MSE of the total estimate with untrimmed weights, which is the estimated variance of the untrimmed total estimate for *cash on hand*. At truncation points above five times the mean, the MSEs are all roughly equal to the MSE for the untrimmed

weights, and so trimming at five times the mean does not create a large difference in the MSE.

In the *ipfraking* command, we set the tolerance between raked and control totals to 0.01 and the procedure converges in 11 iterations. The diagnostic plots (Figure A4 bottom panel) show that the discrepancies between weighted totals and control totals decrease rapidly and change very little beyond the fourth or fifth iteration.

To test for robustness, we also rake with the tolerance at 0.001. In this case, the number of iterations increases to 25. The weights change slightly but remain very similar to the weights with tolerance 0.01, given that the correlation between the two sets of weights is about 1.00. The high correlation can be seen in Figure A5, a scatter plot of the weights generated with tolerance 0.01 and 0.001. A scatter plot of the weights generated by Stata and those generated by R is shown in Figure A6. They are not as highly correlated because of the different ways in which trimming is performed in R and Stata (see Table A1).

Figure A7 shows the command used to produce the raked weights for the full data set, which can be divided into three subsamples based on method of recruitment. The rationale for raking after collapsing the subsamples into a single sample is offered in Appendix B. For more details on the raking procedure, see Vincent (2015).

Step 3: Generate the replicate raking weights using the bootstrap

We generate the replicate raking weights by calling *ipfraking* after *bsweights* (Kolenikov 2010). Iterative proportional fitting is therefore performed on each bootstrapped sample after all the samples have been generated.

As shown in Figure A8, we specify the strata and initial post-stratification weights before running *bsweights* so that the resampling procedure mimics stratified random sampling, using the same strata as the original survey design.

Figure A9 shows the *bsweights* command used to create 300 sets of replicate weights, with the seed set to 2014. We choose to create 300 replicates based on Figure A10, which shows the variance estimates of the mean of the variable *cash on hand* varying with respect to different numbers of bootstrap replications. Twenty trials were conducted for each replication size from 50 to 1,000 in intervals of fifty, and box plots are produced of the resulting variance estimates. The variance estimates are stable with respect to the number of replications, with the median estimates all approximately 13. Moreover, Figure A11 shows that with 300 replications, the average time to compute the variance is roughly 3 seconds.

Step 4: Declare the bootstrap survey environment in Stata

In the declaration of survey environment, we specify the bootstrap option (*vce[bootstrap]*) and the names of the replicate weights (*bsrw[bsw_final_1-bsw_final_300]*) so that variances will be calculated using the replicate raking weights. We then estimate the total value of the *cash on hand* variable and its variance under the bootstrap environment (see Figure A12).

Figure A1: Estimating total value of *cash on hand* with strata under design weights

```

r; t=0.05 14:02:45
. gen simplewgt = 26502410/3663
r; t=0.00 14:02:45
. svyset [pw=simplewgt], strata(strata)
      pweight: simplewgt
      VCE: linearized
Single unit: missing
Strata 1: strata
  SU 1: <observations>
  FPC 1: <zero>
r; t=0.02 14:02:45
. svy: total cashonhand_ed
(running total on estimation sample)

Survey: Total estimation
Number of strata =      84      Number of obs   =      3651
Number of PSUs   =     3651      Population size = 26415588
                                   Design df       =      3567

```

	Total	Linearized Std. Err.	[95% Conf. Interval]	
cashonhand_ed	2.17e+09	9.83e+07	1.97e+09	2.36e+09

```

r; t=0.08 14:02:45

```

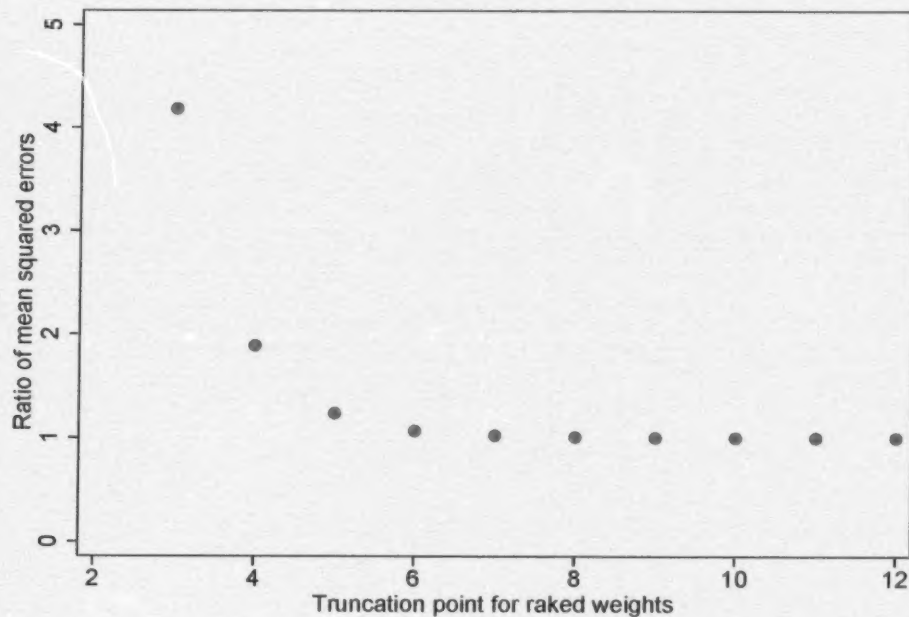
Note: The total of *cash on hand* is estimated with design weights (see Step 1). The number of respondents with non-missing values for *cash on hand* is 3,651.

Figure A2: Setting up a survey environment with post-stratification weights

```
r; t=0.06 14:02:46
. svyset [pw=simplewgt], strata(strata) poststrata(strata) postweight(Count)
      pweight: simplewgt
      VCE: linearized
Poststrata: strata
Postweight: Count
Single unit: missing
Strata 1: strata
      SU 1: <observations>
      FPC 1: <zero>
r; t=0.02 14:02:46
```

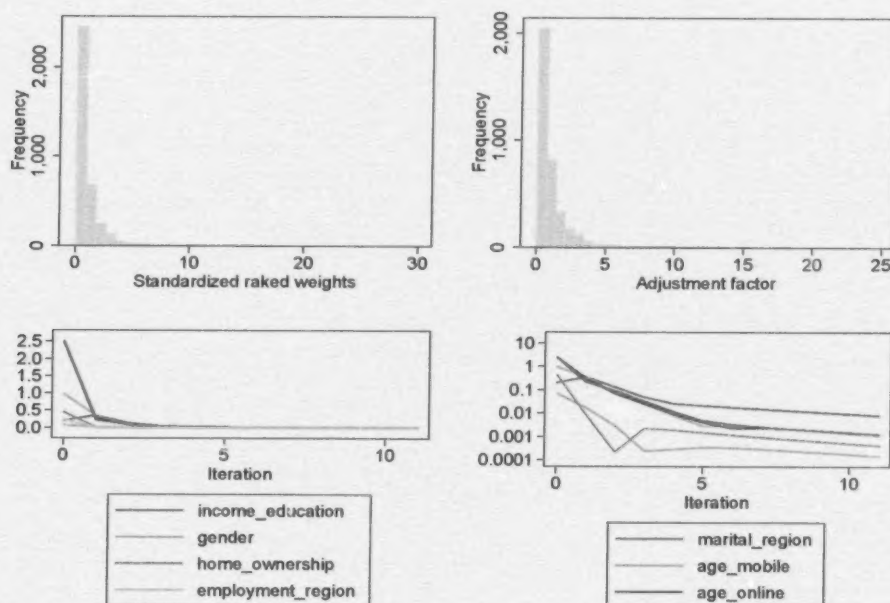
Note: The variable *strata* is the post-strata category and the variable *Count* contains the population size of each post-stratum (see Step 1).

Figure A3: Ratios of mean squared errors for *cash on hand* using trimmed raked weights



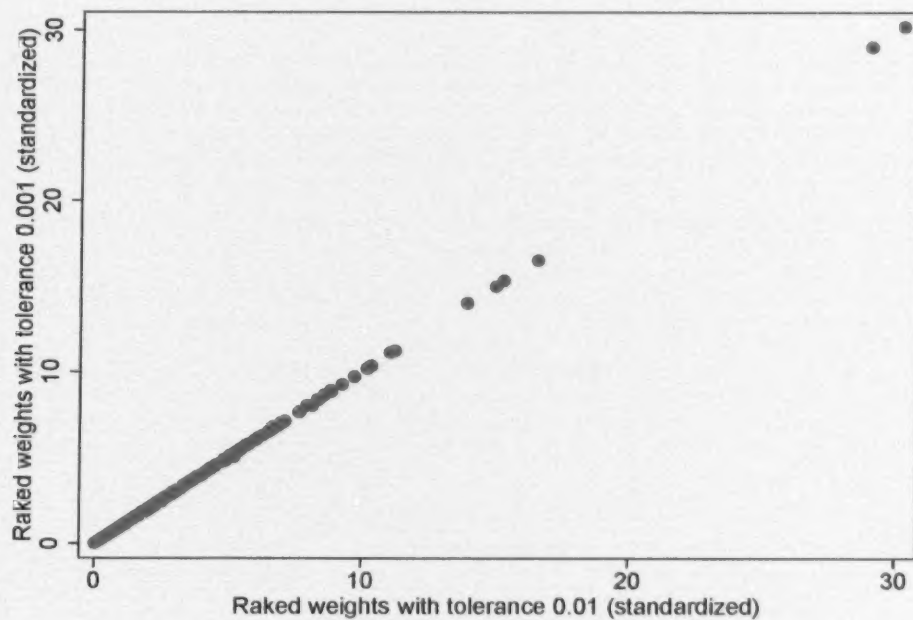
Note: MSEs are calculated for the *cash on hand* variable taking the total computed with untrimmed weights to be the true total (see Step 2). They are then divided by the MSE of the total estimate with untrimmed weights (which equals the variance of the total estimate with untrimmed weights). The resulting ratios are plotted in Figure A3.

Figure A4: Diagnostic graphs for *ipfraking* with no trimming and optimization tolerance 0.01



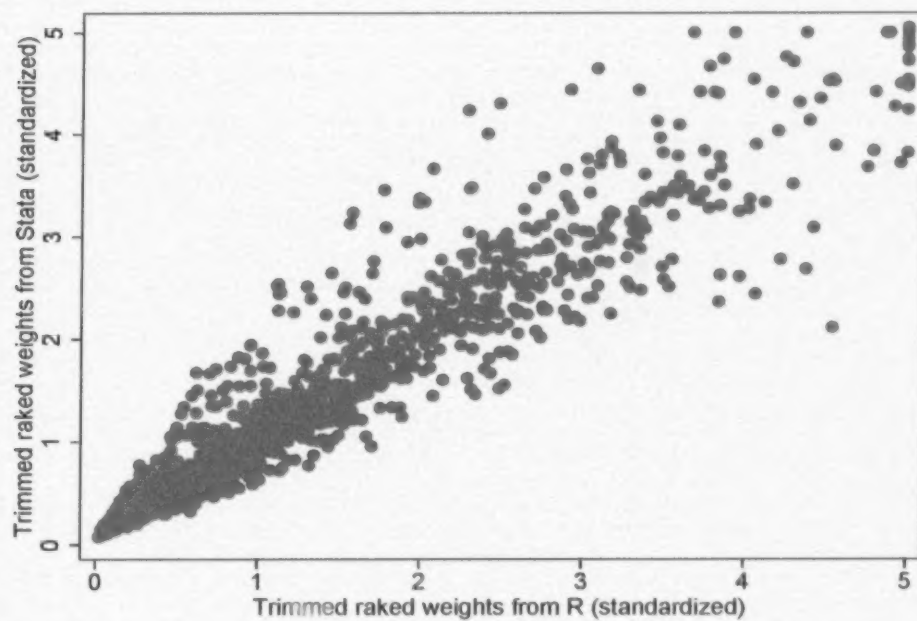
Note: These graphs show the result of performing raking based on CIUS data (Step 2). Starting from the upper left corner and moving clockwise, we have a histogram of the raked weights produced, a histogram of the ratio between the raked weights and initial weights, a plot on a log scale showing the discrepancy with CIUS targets by the number of iterations, and the same plot on an absolute scale. The scale on the x-axis of the histogram in the upper left-hand corner has been standardized by the mean of the raked weights.

Figure A5: Scatter plot of untrimmed raked weights with tolerance 0.01 and tolerance 0.001



Note: Both sets of weights have been standardized by the mean of the untrimmed raked weights with tolerance 0.01 (see Step 2).

Figure A6: Scatter plot of trimmed raked weights generated from Stata and R



Note: Both sets of weights have been standardized by the mean of the untrimmed raked weights generated from Stata with tolerance 0.01 (see Step 2).

Figure A7: The *ipfraking* command

```
r; t=0.06 17:09:32
. ipfraking [pw=$initwgt], generate(rakedwgt) ctotal($rakeva
> rs2) tolerance($tolerance) trace
(output omitted)
r; t=7.68 17:09:40
. summarize rakedwgt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rakedwgt	3663	7659.569	11336.17	153.4302	231592.8

```
r; t=0.01 17:09:40
. global nmean = $trim * r(mean)
r; t=0.00 17:09:40
. ipfraking [pw=$initwgt], generate(rakedwgt_tr) ctotal($rakev
> rs2) tolerance($tolerance) trimhiabs($nmean) trimfrequency($trimfreq) nograph
(output omitted)
r; t=1.14 17:09:41
```

Note: The *ipfraking* command is used twice because we use it to generate the untrimmed weights first. We then use five times the mean of the untrimmed weights as the upper bound for trimming the second time we use the *ipfraking* command. This produces the trimmed weights (see Step 2).

Figure A8: Specifying initial survey environment for *bsweights*

```
r; t=0.06 13:27:36  
. svyset [pweight=$initwgt], strata(strata) psu(id)  
    pweight: basewgt  
    VCE: linearized  
    Single unit: missing  
    Strata 1: strata  
    SU 1: id  
    FPC 1: <zero>  
r; t=0.00 13:27:36
```

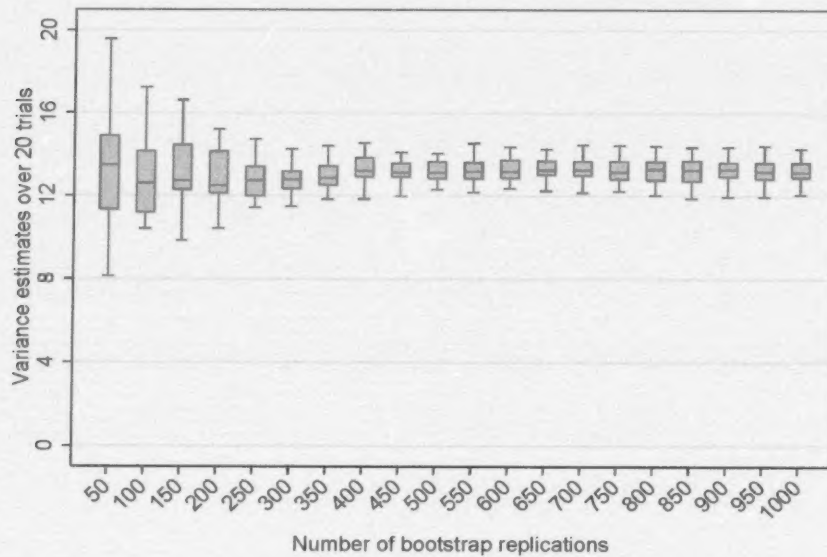
Note: *initwgt* is a placeholder for the initial weight to be used for generating the bootstrap replicate weights (see Step 3). In this case, it is the post-stratification weights.

Figure A9: The *bsweights* command

```
r; t=0.05 13:27:36
. set seed 2014
r; t=0.00 13:27:36
. bsweights bsw_, reps(300) n(-1) dots
Warning: the first-order balance was not achieved
(output omitted)
r; t=31.92 13:28:08
. forvalues i=1/300{
2.     ipfraking [pw=bsw_`i'], generate(bsw_untr_`i') nograph
> ctotol($rakevars) tolerance($tolerance)
3.     summarize bsw_untr_`i'
4.     global nmean = $trim * r(mean)
5.     ipfraking [pw=bsw_`i'], generate(bsw_final_`i') nograph
> ctotol($rakevars) tolerance($tolerance) trimhiabs($nmean) trimfrequency($trimfreq)
6. }
```

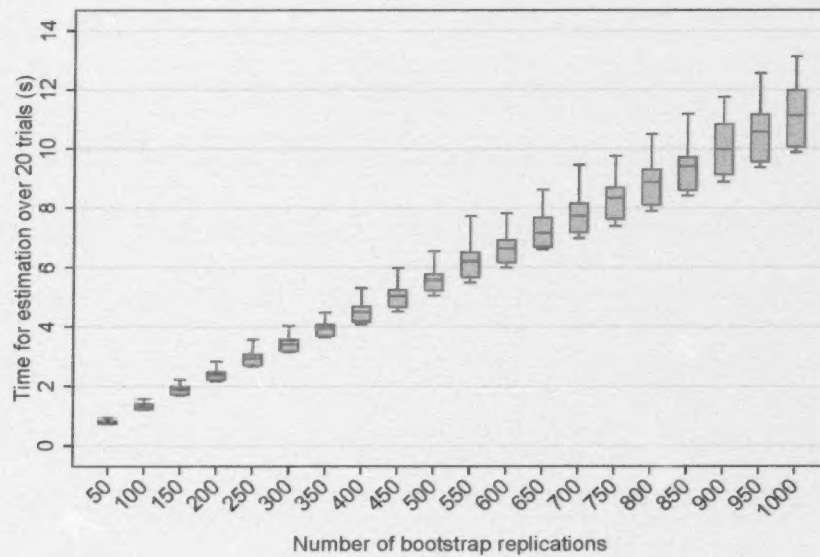
Note: We first produce bootstrap replicate weights using the post-stratification weights as the initial weights. We then calibrate each of the 300 sets of bootstrap replicate weights using the same raking procedure used for the complete sample (Step 3).

Figure A10: Box plots of estimated variance for the mean of *cash on hand*



Note: 20 variance estimates by the bootstrap resampling method (using different seeds) are generated for each replication size from 50 to 1000 in intervals of fifty (see Step 3). Box plots of the 20 estimates are shown for each replication size. Outliers are not shown in the box plots.

Figure A11: Number of bootstrap replications versus time to compute variance



Note: 20 variance estimates by the bootstrap resampling method (using different seeds) are generated for each replication size from 50 to 1000 in intervals of fifty (see Step 3). For each replication size, a box plots is shown of the lengths of time taken for the 20 trials. Outliers are not shown in the box plots.

Figure A12: Estimating total of *cash on hand* with replicate weights

```
r; t=0.05 14:02:46
. svyset [pw=rakedwgt_tr], vce(bootstrap) bsrw(bsw_final_1-bsw
> _final_300) dof(3570)
```

(output omitted)

```
r; t=0.02 14:02:46
```

```
. svy: total cashonhand_ed
(running total on estimation sample)
```

Bootstrap replications (300)

```
-----|-----|-----|-----|-----|-----|
| 1 | 2 | 3 | 4 | 5 |
|-----|-----|-----|-----|-----|
..... 50
..... 100
..... 150
..... 200
..... 250
..... 300
```

```
Survey: Total estimation      Number of obs   =      3651
                             Population size   =  26153773
                             Replications      =       300
                             Design df        =      3570
```

	Observed Total	Bootstrap Std. Err.	Normal-based [95% Conf. Interval]	
cashonhand_ed	2.21e+09	7.95e+07	2.06e+09	2.37e+09

```
r; t=2.89 14:02:49
```

Note: The total of *cash on hand* is estimated using raked weights and the variance is estimated with the bootstrap replicate weights (see Step 4). The number of respondents with non-missing values for *cash on hand* is 3,651.

Table A1: Comparison of the raking commands in R and Stata

	<i>Rake</i> command in R	<i>ipfraking</i> command in Stata
Convergence criteria	Based on the following metric between successive iterative weights: $\max_{j \in S} w_j^{k,p} / w_j^{k-1,p} - 1 $, where $w_j^{k,p}$ is the j th unit weight at the k th step. Convergence is declared when the metric falls below a user-specified tolerance level (Lumley 2012).	Similarly based on the iterative weights but also allows user to specify a tolerance level for matching margins. This tolerance level does not affect convergence, but Stata issues a warning message if it is not satisfied.
Numerical solution	Uses the Newton-Raphson method with analytical derivatives (Lumley 2012).	Uses the Newton-Raphson method with numerical derivatives.
The order of constraints	Does not matter if process converges.	Does not matter if process converges.
Trimming	Different options available: trimming after each iteration or trimming at end of raking process. Does not redistribute weights to untrimmed observations.	Trims at the end of the raking process and redistributes trimmed weights to untrimmed observations.

Note: The *Rake* command is from the R package *survey* written by Thomas Lumley. The *ipfraking* command is from the Stata package *ipfraking* by Stanislav Kolenikov.

Appendix B Rationale for Collapsing Three Subsamples

The full 2013 MOP survey sample consists of three distinct subsamples: online (respondents selected from an online panel of potential volunteers and invited by email)⁹, offline CFM (recent participants of the Canadian Financial Monitor survey invited by regular mail), and Non-CFM (respondents selected from an offline panel of potential volunteers and invited by regular mail). In this section, we justify performing calibration on the full combined data set instead of on the three subsamples separately.

Step 1: We compare both the missing patterns and the distributions of the completed SQ questions for the online, offline CFM, and Non-CFM subsamples, and we find the following:

- For the 2013 MOP survey, the offline CFM and offline Non-CFM respondents have lower item-response rates than the online respondents. However, among offline respondents, CFM and Non-CFM respondents have similar missing data patterns.
- Regarding the distributions of answers to the survey questions, we find significant differences between the online and offline respondents. Nevertheless, there are few differences between the offline CFM and offline Non-CFM. For the completed survey questions, we conduct the two-sample omnibus test introduced by Epps and Singleton, which usually has greater power and is suitable for both discrete and continuous valued data. However, the Epps-Singleton test may fail when data are heavily concentrated in a single value (almost degenerating to a constant), in which case we conduct the Mann-Whitney rank-sum (MW) test instead.

Action: Combine the offline CFM and Non-CFM to generate the paper-based frame (to achieve a larger sample size and better MSE performance), while keeping the online subsample as a separate frame.

Step 2: Perform the analysis suggested by Brick et al. (2006) and Young et al. (2012) to compare the means and variances across different weighting schemes (see below).

- Based on the above discussion, we determine that we have a classical dual-frame sample. In particular, we have a non-overlapping dual-frame sample because panelists can only be invited once for the MOP survey, although they are allowed to be in both the paper-based and online subsamples. Ipsos Reid sent invites to the paper-based frame first, then

⁹ Probability sampling ensures that every possible sample from the population has a known probability of being chosen (Lohr 2010). Because the online panel is not recruited using probability sampling, Bayesian credibility intervals are usually calculated for estimates based on this panel (Roshwalb, El-Dash and Young 2012). However, in this report, frequentist variance estimation is used to measure the sampling errors for the pseudo-population (Quatember 2016) instead of for the true population.

de-duplicated those who had been invited from the paper-based when sending the online invites. The advantage of a non-overlapping dual-frame sample is that we have a simple way to collapse different frames and do not need to adjust for the composite weights as in an overlapping dual-frame sample (Callegaro et al. 2011).

- **Option 1** Only use the design weights without the raking procedure: Obtain design weights for the paper-based frame as W_P^D and the online frame as W_O^D , respectively. Then merge the two frames with the weights $\{W_P^D, W_O^D\}$. According to Vincent (2015), we will produce two sets of design weights, which can be fed into the raking procedure.
- **Option 2** Merge and then rake: Rake the combined data set with the initial weights $\{W_P^D, W_O^D\}$, and this will generate the raked weights $\{W_P^R, W_O^R\}$. A similar method was implemented for two surveys: (1) Canadian Community Health Survey and (2) the Survey Data on the Health of New Yorkers.
- **Option 3** Rake separately for each frame, then merge: Rake each frame separately, with the initial weights from Option 1. This will produce the raked weights $\{W_P^{Rr}\}$ and $\{W_O^{Rr}\}$ for the paper-based and online frames. Lastly we merge $\{W_P^{Rr}\}$ and $\{W_O^{Rr}\}$ to generate the merged weights $\{W_P^{Rr}, W_O^{Rr}\}$.
- **Option 4** Merge and then rake with an extra Internet access control variable indicating whether the panelist has access to the Internet at home. Rake the combined data set with the initial weights $\{W_P^D, W_O^D\}$ with an extra Internet access control variable, and this will generate the raked weights $\{W_P^{RI}, W_O^{RI}\}$.
- **Option 5** Only rake the paper-based frame, producing the raked weights $\{W_P^{Rr}\}$.

Table B1 shows the means and variances of different weighting schemes with uniform weights as the initial weights for raking, while Table B2 is constructed similarly except using post-stratification weights for raking. In terms of weighted mean estimates, differences across weighting schemes are less than five percent. The differences between the estimated means are similar even with different initial weights. As for the variances associated with different weighting schemes, the variance from Option 2 is the second smallest in Table B1 and the smallest in Table B2, suggesting that using the raking in the last step is able to reduce the variance dramatically (D'Arrigo and Skinner 2010).

Action: Choose Option 2, which first merges the two frames and then applies the raking as for the calibration of a full data set.

Table B1: Means and variances of *cash on hand* using uniform weights in raking

	Option 1 (N/n)	Option 2	Option 3	Option 4	Option 5
Mean	81.96	77.95	76.94	80.10	79.24
Variance	14.72	7.02	26.08	9.22	6.34

Note: Variance of Option 1 is calculated based on the linearization method, which is the default in Stata, while variance estimates from Options 2 to 5 are computed based on the bootstrap resampling.

Table B2: Means and variances of *cash on hand* using post-stratification weights in raking

	Option 1 (N_h/n_h)	Option 2	Option 3	Option 4	Option 5
Mean	81.91	78.27	76.91	79.94	82.25
Variance	16.76	7.13	56.27	9.33	11.78

Note: Variance of Option 1 is calculated based on the linearization method, which is the default in Stata, while variance estimates from Options 2 to 5 are computed based on the bootstrap resampling.